

Science
June 11, 2004

**EDUCATION RESEARCH:
Meager Evaluations Make It Hard to Find Out What Works
by Jeffrey Mervis**

A popular buzzword in U.S. education these days is discovering "what works." The Education Department even funds a "What Works Clearinghouse" on programs ranging from teaching math to reducing schoolyard violence. This heightened interest in assessment stems from the massive 2001 education reform bill--known as the No Child Left Behind Act--which requires school districts to offer programs shown to be effective through "scientifically based research." But there's a dirty little secret behind that requirement: No program has yet met that rigorous standard, because none has been scientifically evaluated and shown to be effective. (A related secret is that there's no consensus on the type of evaluation studies that are needed.)

Two reports issued last month highlight the problem--and suggest ways to bolster the embryonic field of evaluation research. One, by the National Research Council (NRC), examined evaluations of 19 elementary and secondary school mathematics curricula and found them wanting.

<http://www.sciencemag.org/cgi/content/full/304/5677/1583#affiliation#affiliation> > *

The second, from a public-private consortium known as Building Engineering and Science Talent (BEST), did the same for programs aimed at increasing the number of minorities, women, and low-income students studying science and math (bestworkforce.org). Again, none of the programs could claim to be successful based on objective assessments.

"The sad reality is that we really don't know what works. And that leaves people in the lurch," says William Schmidt, a member of the NRC panel and a mathematics educator at Michigan State University in East Lansing. Schmidt, who is also U.S. coordinator for an ongoing study comparing student math and science achievement in 40-plus countries, says that conducting high-quality evaluations is possible, "but it takes a real national commitment. And money."

Both reports emphasize that many programs may be doing a terrific job of helping children. It's just that there's no way to tell, scientifically. "Evidence matters, ... [but] reliable empirical evidence is hard to find," notes the BEST report, which compiled material on some 200 programs before deciding that only 34 had been studied with sufficient rigor to justify further scrutiny. Of the 20 that it was able to rate, none met BEST's gold standard: five studies conducted by independent evaluators that showed substantially positive results.

The NRC study took a different approach but reached similar conclusions. Its survey of the literature identified 698 studies on 13 math curricula developed by the National

Science Foundation (NSF) and six by commercial publishers. But only 147 (21%) met the panel's criteria for weighing effectiveness. It divided those studies into four types of evaluations before concluding that "the corpus of studies does not permit one to determine the effectiveness of individual programs." In other words, there weren't enough good data to draw definitive conclusions about their value in the classroom.

The most surprising message from the two studies may be that evaluation experts aren't surprised by the results. "The entire discipline of rigorous evaluation is just emerging," says Judith Ramaley, head of the education directorate at NSF, which funded the NRC study and helped launch BEST in 2001. "But the good news is that the study provides us with a wonderful user's manual for how to go about our business. And it comes at a time when we are finally capable of doing this type of evaluation."

Before that happens, however, experts need to agree on what constitutes a rigorous evaluation. The NRC panel devoted most of its 212-page report to that topic. The problem is complicated by the many factors that influence student achievement: students' previous knowledge, their teachers' quality of training, the level of resources available, the degree of parental and community support, and so on. "This has never been done before," says panel chair Jere Confrey of Washington University in St. Louis, Missouri. "People may think that it's a simple problem, but it's really quite difficult." The increased reliance on tests to determine the fate of students and their schools has increased pressure on educators and evaluators to get it right, she notes.

Evaluation researchers are moving slowly toward the type of evidence-based standards used in biomedical research, says Carlos Rodriguez of the American Institutes for Research, a Washington, D.C.-based nonprofit that worked with BEST to define principles for both designing an effective program and assessing its performance. But there are limits. "NSF is now asking for multivariate and controlled studies," he says. "But you have to remember that we are measuring human behavior, and that's hard to quantify."

Controlled studies are especially difficult in an educational setting, points out Mary Catherine Swanson, whose Advancement via Individual Determination (avidonline.org <<http://avidonline.org/>>), a national program for at-risk students, was praised by BEST for "notable effectiveness" but faulted for the absence of studies involving students not in the program. "It's hard to exclude people from a program that they think is working," she says. Swanson thinks that BEST is correct in setting the bar high, however, and she hopes that a recent expansion of the 24-year-old program to Canada--where educators plan to follow students eligible for AVID but unable to work it into their schedules--will provide clearer signs of the program's effectiveness.

Once more rigorous evaluations are in place, says Rodriguez, educators and the public must be willing to resist the temptation to oversimplify the results. "There are no magic bullets," he says. "We want to find what works most of the time for most students. But we still have to adapt programs to fit the population being served." From last Friday's issue of Science:

EDUCATION RESEARCH:

Meager Evaluations Make It Hard to Find Out What Works

Jeffrey Mervis

A popular buzzword in U.S. education these days is discovering "what works." The Education Department even funds a "What Works Clearinghouse" on programs ranging from teaching math to reducing schoolyard violence. This heightened interest in assessment stems from the massive 2001 education reform bill--known as the No Child Left Behind Act--which requires school districts to offer programs shown to be effective through "scientifically based research." But there's a dirty little secret behind that requirement: No program has yet met that rigorous standard, because none has been scientifically evaluated and shown to be effective. (A related secret is that there's no consensus on the type of evaluation studies that are needed.)

Two reports issued last month highlight the problem--and suggest ways to bolster the embryonic field of evaluation research. One, by the National Research Council (NRC), examined evaluations of 19 elementary and secondary school mathematics curricula and found them wanting.

<http://www.sciencemag.org/cgi/content/full/304/5677/1583#affiliation#affiliation> *

The second, from a public-private consortium known as Building Engineering and Science Talent (BEST), did the same for programs aimed at increasing the number of minorities, women, and low-income students studying science and math (bestworkforce.org). Again, none of the programs could claim to be successful based on objective assessments.

"The sad reality is that we really don't know what works. And that leaves people in the lurch," says William Schmidt, a member of the NRC panel and a mathematics educator at Michigan State University in East Lansing. Schmidt, who is also U.S. coordinator for an ongoing study comparing student math and science achievement in 40-plus countries, says that conducting high-quality evaluations is possible, "but it takes a real national commitment. And money."

Both reports emphasize that many programs may be doing a terrific job of helping children. It's just that there's no way to tell, scientifically. "Evidence matters, ... [but] reliable empirical evidence is hard to find," notes the BEST report, which compiled material on some 200 programs before deciding that only 34 had been studied with sufficient rigor to justify further scrutiny. Of the 20 that it was able to rate, none met BEST's gold standard: five studies conducted by independent evaluators that showed substantially positive results.

The NRC study took a different approach but reached similar conclusions. Its survey of the literature identified 698 studies on 13 math curricula developed by the National Science Foundation (NSF) and six by commercial publishers. But only 147 (21%) met

the panel's criteria for weighing effectiveness. It divided those studies into four types of evaluations before concluding that "the corpus of studies does not permit one to determine the effectiveness of individual programs." In other words, there weren't enough good data to draw definitive conclusions about their value in the classroom.

The most surprising message from the two studies may be that evaluation experts aren't surprised by the results. "The entire discipline of rigorous evaluation is just emerging," says Judith Ramaley, head of the education directorate at NSF, which funded the NRC study and helped launch BEST in 2001. "But the good news is that the study provides us with a wonderful user's manual for how to go about our business. And it comes at a time when we are finally capable of doing this type of evaluation."

Before that happens, however, experts need to agree on what constitutes a rigorous evaluation. The NRC panel devoted most of its 212-page report to that topic. The problem is complicated by the many factors that influence student achievement: students' previous knowledge, their teachers' quality of training, the level of resources available, the degree of parental and community support, and so on. "This has never been done before," says panel chair Jere Confrey of Washington University in St. Louis, Missouri. "People may think that it's a simple problem, but it's really quite difficult." The increased reliance on tests to determine the fate of students and their schools has increased pressure on educators and evaluators to get it right, she notes.

Evaluation researchers are moving slowly toward the type of evidence-based standards used in biomedical research, says Carlos Rodriguez of the American Institutes for Research, a Washington, D.C.-based nonprofit that worked with BEST to define principles for both designing an effective program and assessing its performance. But there are limits. "NSF is now asking for multivariate and controlled studies," he says. "But you have to remember that we are measuring human behavior, and that's hard to quantify."

Controlled studies are especially difficult in an educational setting, points out Mary Catherine Swanson, whose Advancement via Individual Determination (avidonline.org <<http://avidonline.org/>>), a national program for at-risk students, was praised by BEST for "notable effectiveness" but faulted for the absence of studies involving students not in the program. "It's hard to exclude people from a program that they think is working," she says. Swanson thinks that BEST is correct in setting the bar high, however, and she hopes that a recent expansion of the 24-year-old program to Canada--where educators plan to follow students eligible for AVID but unable to work it into their schedules--will provide clearer signs of the program's effectiveness.

Once more rigorous evaluations are in place, says Rodriguez, educators and the public must be willing to resist the temptation to oversimplify the results. "There are no magic bullets," he says. "We want to find what works most of the time for most students. But we still have to adapt programs to fit the population being served."

